2012

# Improved Speech Intelligibility with a Chimaera Hearing Aid Algorithm

Andrew Hines
*Technological University Dublin*, andrew.hines@tudublin.ie

Naomi Harte
*University of Dublin, Trinity College*



## Recommended Citation

Hines, A. & Harte, N. (2012) Improved Speech Intelligibility with a Chimaera Hearing Aid Algorithm, *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*Portland, Oregon, USA, September 9-13.

# Improved Speech Intelligibility with a Chimaera Hearing Aid Algorithm

*Andrew Hines*[1]*, Naomi Harte*

Sigmedia Group, Department of Electronic & Electrical Engineering
Trinity College Dublin, Ireland
andrew.hines@tcd.ie[1]

## Abstract

It is recognised that current hearing aid fitting algorithms can corrupt fine timing cues in speech. This paper presents a fitting algorithm that aims to improve speech intelligibility, while preserving the temporal fine structure. The algorithm combines the signal envelope amplification from a standard hearing aid fitting algorithm with the fine timing information available to unaided listeners. The proposed "chimaera aid" is evaluated with computer simulated listener tests to measure its speech intelligibility for 3 sample hearing losses. In addition, the experiment demonstrates the potential application of auditory nerve models in the development of new hearing aid algorithm designs using the previously developed Neurogram Similarity Index Measure (NSIM) to predict speech intelligibility. The results predict that the new aid restores envelope without degrading fine timing information.

**Index Terms**: auditory periphery model, NSIM, speech intelligibility, Hearing Aid

## 1. Introduction

Developing new hearing aids algorithms is time and labour intensive. Each new algorithm needs a large test-set covering a range of hearing losses to properly evaluate their potential. Prior work by the authors showed that a computational model of the auditory periphery allowed simulated listener tests, where real listeners were substituted with a computer model. Speech intelligibility performance for normal and hearing impaired listeners, as well as listeners with hearing aids, were shown to be predicted using a novel Neurogram Similarity Index Measure (NSIM) [1, 2]. The intelligibility level is predicted by comparing internal representations (neurograms) of speech sounds that visually represent the neural discharge activity in auditory nerve fibres spectro-temporally. NSIM has been tested with normal hearing and hearing impaired listeners in a range of conditions (quiet and noise) and shown to predict intelligibility over a range of sound intensity levels [2].

The NAL-RP hearing aid formula aims to maximise intelligibility by making a listener perceive all frequency bands to have equal loudness [3]. It amplifies the envelope of the speech signal at the expense of corrupting fine timing changes that contain important cues. This work presents a novel hearing aid fitting method that aims to improve the temporal fine structure information available for aided hearing impaired listeners. The amplified envelope and original fine timing signals are combined and the resulting speech intelligibility levels assessed.

Section 2 introduces listener test simulation using the auditory nerve model and NSIM. Temporal fine structure and auditory chimaeras are also discussed. Section 3 describes the simulation methodology including the hearing profiles used and chimaera hearing aid fitting method. Section 4 presents the simulated results and Section 5 discusses the potential improvement in speech intelligibility using this chimaera aid approach.

## 2. Background

### 2.1. Neurogram Assessment

A neurogram is analogous to a spectrogram. It presents a pictorial representation of a signal in the time-frequency domain using colour to indicate the intensity of neural firing activity. Example neurograms can be seen in Fig. 4.

Speech signals, specifically consonant-vowel-consonant (CVC) words, are presented as inputs to the Zilany et al. auditory nerve (AN) model [4] which simulates the middle and inner ear, and produces simulated AN discharges in response to the signal. Two types of neurograms are assessed: temporal fine structure (TFS) which retained spike timing information; and average discharge rate or temporal envelope (ENV).

The output neural activity is binned into time bins (of 10 $\mu$s and 100 $\mu$s for TFS and ENV) to create post stimulus time histograms (PSTHs) that are then used to create neurograms by convolving them with 50% overlap, 32 and 128 sample Hamming windows respectively. As in prior work [1], neurograms with 30 characteristic frequencies (CFs) are used, spaced logarithmically between 250 and 8000 Hz. The neural response at each CF are created from the PSTH of 50 simulated AN fibres with varying spontaneous rates.

Neurograms for each phoneme of the CVC words are assessed using NSIM. NSIM was adapted from SSIM [5], an image comparison metric, and is used to compare degraded neurograms with a reference neurogram from a normal hearing AN model for the same input signal. It has been shown to be superior to other simple point to point measures such as a relative mean squared error assessed per neurogram element. It is a bounded measure yielding a similarity in the range of 1 for an exact match to 0 for no similarity. NSIM and its use is described in detail in prior work [1].

### 2.2. Temporal fine structure

The structure of speech signals can be segmented by frequency where envelope (ENV) is defined as signal fluctuations between around 2-500 Hz and temporal fine structure (TFS) as signal cues with dominant fluctuations from about 600 Hz - 10 kHz [6, 7, 8].

According to Rosen [6], ENV cues are mainly manner and voicing while TFS are place cues and, to a lesser degree, voicing and nasality. Sheft et al. [9] agreed but found a stronger contribution of TFS cues for voicing than place. Wang et al. [10] assert that ENV is critical for speech perception, whereas TFS is critical for pitch perception. Lorenzi et al. [8] showed that

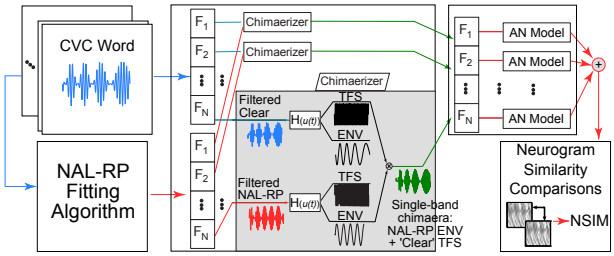


Figure 1: The chimaeriser takes the original signal and the NAL-RP aided signal as inputs and provides a chimaera signal output to the AN model. The AN model, simulating at 30 characteristic frequencies, produces PSTH output that is used to create a neurogram. An NSIM comparison is carried out on neurograms for each of the 150 test phonemes.

while TFS contains cues for speech identification, subjects with a *moderate* loss performed almost as well as normal hearing listeners with both unprocessed and ENV only speech. Under both conditions, normal hearing (NH) and hearing impaired (HI) listeners scored 80-100 %. However, HI listeners struggled with TFS speech scoring less than 20%, while NH listeners remained at around 90% discrimination.

The problems experienced by HI listeners in background noise over and above the issues experienced by NH listeners has been attributed to loss of TFS discrimination by a number of studies [11, 8, 12]. The neural correlates to acoustic ENV and TFS involve looking at the average discharge rate and spike timing information along the auditory nerve. TFS cues are observed in neurograms as the synchronization of the AN spikes phase-locking to the stimulus. Miller et al. [13] champion the use of temporal response pattern analysis as a tool for studying AN information representation as "temporal analysis reveals much about the nature of impairment and provides guidance as to the problems that need to be solved in order to compensate for the impairment". Sheft et al. [9] also suggest that the fidelity of TFS transmission should be measured quantitatively in hearing device assessment.

Bruce et al. [14] investigated the performance of hearing aid methods and found that the TFS neurograms were closer in similarity to a reference neurogram when gain adjustments were below the prescribed gains. It was suggested that spread of synchrony and the change in phase-frequency responses in an impaired ear could be factors but it was left as an open question requiring further investigation.

### 2.3. Auditory Chimaeras

A novel technique to investigate auditory perception using chimaeric sounds was developed by Smith et al. [7]. "Auditory chimaeras" allow the perceptual importance of envelope and fine structure portions of signals to be separated and evaluated. Two input sounds are split through an N band filterbank. The matching band signals are then passed through a chimaerizer, which splits the signal into ENV (the magnitude of the signal) and TFS (the instantaneous phase) using a Hilbert transform. The ENV from the first signal is combined with the TFS from the second signal to produce a single band chimaera that is then summed over all N bands to produce a multiband chimaera. This is illustrated in the shaded area of Fig. 1. Smith et al. carried out a number of tests on speech reception, melody recognition and sound localisation, using chimaeras generated with two different signals comprising of speech-speech, speech-noise and melody-melody signals.

Ibrahim and Bruce [15] reproduced the chimaera results of Smith et al. using the AN model, showing it could be used to predict ENV and TFS speech reception using speech-noise chimaeras over a varied number of chimaeriser frequency bands.

An alternative application of auditory chimaeras by Liu et al. inspired this work [16]. Chimaeras were created from clear and conversational versions of the same speech. "Clear speech" differs acoustically from everyday "conversational speech" in a number of ways, e.g. it includes a slower speech rate, enhanced fundamental frequency variation, expanded vowel space and higher energy distribution. It has been shown to produce high intelligibility scores for tests on normal hearing and hearing impaired listeners, in quiet and in noise. Liu et al. created auditory chimaeras of matched clear and conversational speech using nonuniform stretching to align segments. They found that the clear speech ENV and conversational TFS produced better results in high SNR situations, while the reverse was true in low SNR environments.

This work used the same paradigm as Liu et al. except instead of augmenting intelligibility with "clear speech", the TFS from the original speech signal is combined with the amplified ENV signal from the NAL-RP hearing aid to create a *chimaera aid* to improve intelligibility without corrupting important TFS cues.

## 3. Method

This experiment implemented the system depicted in Fig. 1, creating an auditory chimaera with unprocessed, clear TFS, and NAL-RP aided ENV. The aim of the *chimaera aid* was to provide the listener with aided gain in the ENV portion of the signal but to maintain the TFS fidelity by restoring the original signal's TFS to the signal processed by the hearing aid. The test looked at whether NSIM measurements using the AN model predicted improved TFS neurogram similarity for a range of HI listeners.

Smith et al.'s auditory chimaera algorithm was used to create a chimaera signal based on the envelope of the NAL-RP adjusted signal and the fine structure of the original signal. Three hearing losses (shown in Fig. 2) were simulated with the AN model at presentation level of 55 dB SPL. This level was chosen as it was a level at which the *mild* loss was above its speech reception threshold, while the *moderate* and *severe* were below their threshold unaided but above when aided. Fifty test words were presented to the AN model directly to calculate the unaided NSIM scores using ENV and TFS neurograms. The words were then filtered using the NAL-RP hearing aid formula with the prescribed linear gains.

The 50 test words were filtered through the NAL-RP filter and a 30 band "chimaerizer", as illustrated in Fig. 1, and then presented to the AN model. Phoneme NSIM scores were calculated from the neurogram outputs by comparing them against 65 dB SPL reference neurograms, as in prior work [1]. It was necessary to adjust the time alignment to account for the delay introduced by the chimaerizer to ensure accurate phoneme comparisons.

## 4. Results and Discussion

The results for both ENV and TFS neurogram similarity are presented in Fig. 3. For both ENV (shown on the left) and TFS (shown on the right), results are presented for three hearing impairments, as labelled on top. For each hearing profile, results are presented under three conditions across the x-axis: unaided, NAL-RP aided and *chimaera aided*. These results are also bro-

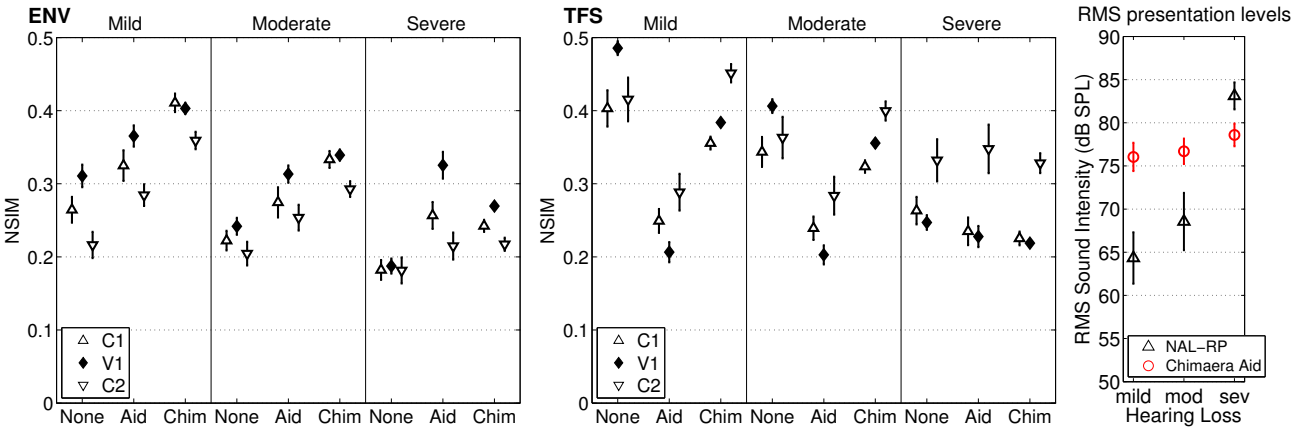


Figure 3: Left and Centre: ENV and TFS NSIM results for unaided (None), NAL-RP aided (Aid) and *chimaera aided* (Chim) listening at 55 dB SPL. The NSIM for each phoneme group (C1,V,C2) are plotted showing +/- 1 s.e. for 3 test losses: *mild*, *moderate* and *severe*. As expected, the ENV results for all phoneme groups predict NAL-RP aided results are higher than the unaided simulations for all hearing losses tested. The *chimaera aided* results mirror this trend. For TFS, the aided simulations score lower than the corresponding unaided results but the *chimaera aid* reverses this trend and maintains the TFS NSIM scores at levels comparable to the unaided simulations. Right: Output intensity levels for the 50 words tested after applying the prescription gains for the 3 test hearing losses.

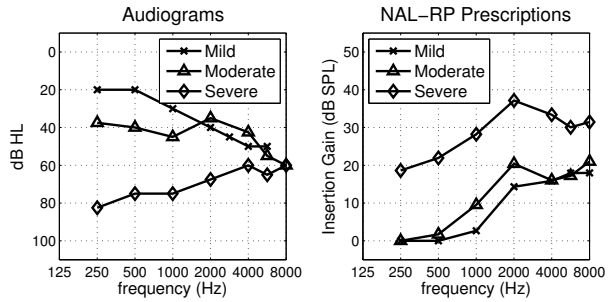


Figure 2: Actual listener audiograms (as used in [2]) and corresponding hearing aid prescriptions for the hearing losses tested.

ken down by phoneme group (C1 △,V1 ♦,C2 ▽).

The ENV results show that, for each phoneme group, the NAL-RP aided NSIM scores are above the unaided scores, predicting that NAL-RP will improve speech intelligibility. Comparing the ENV NSIM aided and chimaera aided results it would be expected that, as the ENV portion of the *chimaera aided* signal had NAL-RP gains applied to it, it should produce comparable NSIM scores. What is actually predicted is an increase in NSIM for the *mild* and *moderate* losses and a decrease for the severe loss. This is likely due to the chimaerizer algorithm used, as prior to recombining the ENV and TFS signal components, the chimaerizer algorithm carries out a peak normalisation on both components. Its impact on the overall gain applied to the signal can be seen in Fig. 3C which shows the root mean squared presentation level of the 50 words tested after applying the prescription gains for the fitting methods. The *chimaera aid* gains are compressed into a smaller range compared to the NAL-RP gains, providing larger gains for *mild* and *moderate* losses but less gain for the *severe* loss. As a result the words are actually presented at levels below threshold for some frequencies in case of the *severe* loss, resulting in poorer ENV NSIM scores.

The TFS results predict for the *chimaera aid* comparable improvements to the regular NAL-RP results. The TFS NSIMs show that, for the *mild* and *moderate* losses, the *chimaera aided* results restore the NSIM scores to the unaided levels, improving them from the floor level of the aided results. In the *severe* loss case, the unaided results are at a comparably low level to the aided results and the chimaera results don't show any significant improvement in neurogram similarity.

The results imply that, for the *severe* loss, the TFS reception has been impaired and cannot be augmented by supplying a clean TFS as the broadened auditory filters are not supplying a quality TFS signal to the auditory nerve. This could be a failure to use higher-frequency speech cues, even though the frequency bands have been made audible by the hearing aid. It was suggested by Hopkins et al. [11] that additional TFS information may not help a severely impaired listener, due to a general problem with higher-frequency speech components. In the moderate case, the unaided TFS results are restored by the chimaera aided signal, suggesting that the user could potentially benefit from the fine timing as well as the envelope intelligibility cues.

The vowel neurograms in Fig. 4 illustrate structural and intensity features that NSIM is capturing in its similarity scores. The unaided ENV neurograms show the lack of spectral cues, with the F0 formant visible for the *mild* loss but nothing for the *moderate* or *severe* loss. The corresponding aided neurograms show that there is information available at higher frequencies, but that the higher formant information has spread to higher frequencies in the case of the *severe* loss. The TFS neurograms illustrate the phase-locking and spread of synchrony for progressively impaired listeners. It should be noted that it is important not to read too much into any specific example's NSIM score. The TFS scores were calculated over a neurogram for the complete vowel, not just the 20 ms snapshot presented. The error bars in the results for tests over 50 phonemes warn against comparing the example scores and judging on one example. A minimum floor threshold in NSIM scores occurs as even the absence of features will be measured as a sign of similarity, e.g. a quiet pause before a plosive burst.

The NAL-RP aided results highlighted the tradeoffs made in corrupting the TFS signal to add sufficient gain in the ENV to ensure that ENV speech cues are available to the hearing impaired listener. These results tie in with the observations made by Bruce et al., that the spike timing information for aided listeners was better as gains decreased rather than increased. The *chimaera aid* is predicted to give the best of both ENV and TFS results for mild to moderate losses. TFS cues for severe losses

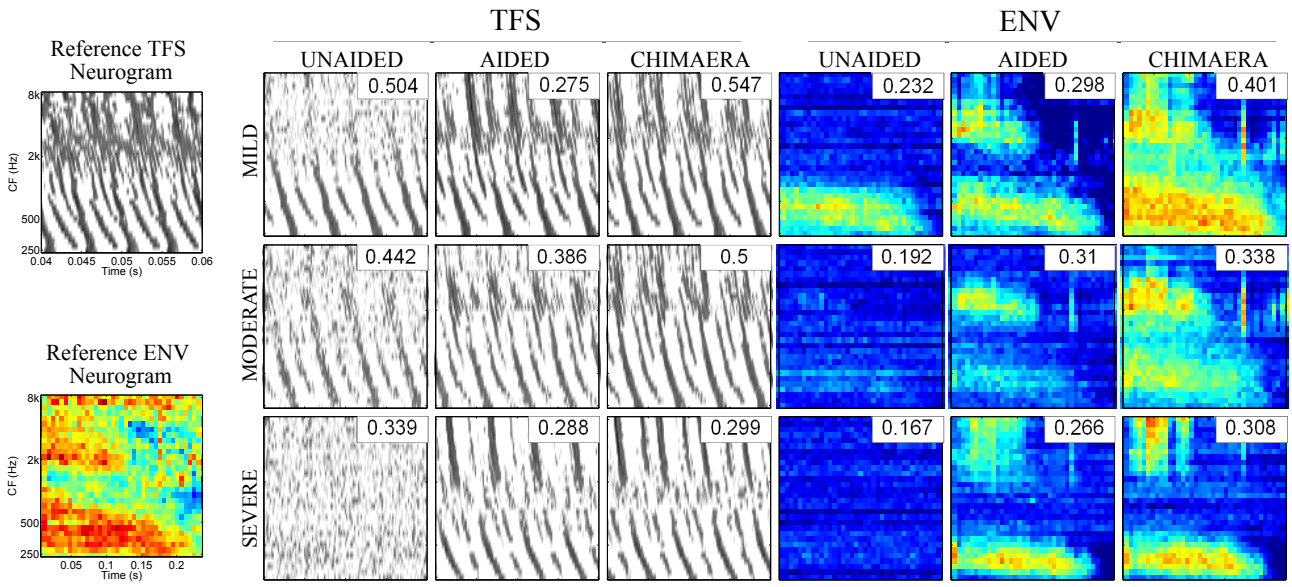


Figure 4: To the left are reference TFS and ENV neurograms for the vowel /ow/ at 55 dB SPL. Neurograms for the same vowel modelled with *mild*, *moderate* and *severe* HIs in unaided, NAL-RP aided and chimaera aided conditions are shown. These illustrate the effect of the different inputs on the ENV and TFS neurograms. The NSIM scores for comparisons against the reference neurograms are shown above each neurogram. The time range covers the full vowel in the ENV neurograms (approx 240 ms) and a snapshot of 20 ms of the vowel starting after 40ms. Axes labels, omitted on the sample results, match those shown on the reference neurograms.

were not restored, as the ability to use TFS was not available irrespective of the presentation level.

These results demonstrate the promising potential of hearing aid design using simulated speech tests. Tests in noise are the obvious next step, as this is where TFS is viewed as being important for speech cues. Tests over a variety of presentation levels could also strengthen the predicted benefits, although Sheft et al. observed that identification of consonants with TFS is robust to variations of stimulus level. Further investigation into the optimal number of frequency bands in chimaerizer for the *chimaera aid* is required as it was shown to be a critical factor for speech intelligibility by Smith et al. The number used here was chosen to match the approximate number of critical bands within the cochlea and also the number of frequency bands used in the simulations with the AN model. Carrying out tests with real listeners would be the final step in validating the predicted benefits of the chimaera aid.

## 5. Conclusions

It was shown that corruption of TFS speech cues can be reduced using a chimaera hearing aid. The simulations predicted that the *chimaera aid* can restore ENV without degrading TFS. The combination of NSIM and the AN model to develop novel hearing aid designs was demonstrated and, subject to further validation with listener tests, proposed as a useful pre-cursor to trials with real hearing impaired listeners.

## 6. References


[1] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Commun.*, vol. 54, no. 2, pp. 306 – 320, 2012.

[2] A. Hines and N. Harte, "Comparing hearing aid algorithm performance using simulated performance intensity functions," in *Speech Perception and Auditory Disorders, Int. Symposium on Audiological and Auditory Research (ISAAR)*, 2011.

[3] H. Dillon, *Hearing Aids*, Thieme Medical Publishers, NY, 2001.

[4] M. S. A. Zilany, I. C. Bruce, P. C. Nelson, and L. H. Carney, "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics," *J Acoust Soc Am*, vol. 126, no. 5, pp. 2390–2412, 2009.

[5] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE T Image Process*, vol. 13, no. 4, pp. 600–612, 2004.

[6] S. Rosen, "Temporal information in speech: Acoustic, auditory and linguistic aspects," *Phil. Trans. R. Soc. B*, vol. 336, no. 1278, pp. 367–373, 1992.

[7] Z.M. Smith, B. Delgutte, and A.J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 6876, pp. 87–90, 2002.

[8] C. Lorenzi, G. Gilbert, and and S. Garnierand B.C.J. Moore H. Carn, "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proc. Nat. Acad. Sci.*, vol. 103, no. 49, pp. 18866–18869, 2006.

[9] S. Sheft, M. Ardoint, and C. Lorenzi, "Speech identification based on temporal fine structure cues," *J Acoust Soc Am*, vol. 124, no. 1, pp. 562–575, 2008.

[10] S. Wang, L. Xu, and R. Mannell, "Relative contributions of temporal envelope and fine structure cues to lexical tone recognition in hearing-impaired listeners," *J. Assoc. Res. Otolaryngol.*, pp. 1–12, 2011.

[11] K. Hopkins, B. C. J. Moore, and M. A. Stone, "Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech," *J Acoust Soc Am*, vol. 123, no. 2, pp. 1140–1153, 2008.

[12] P. Nelson, "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," *J Acoust Soc Am*, vol. 113, no. 2, pp. 961, 2003.

[13] R. L. Miller, J. R. Schilling, K. R. Franck, and E. D. Young, "Effects of acoustic trauma on the representation of the vowel /epsilon/ in cat auditory nerve fibers," *J Acoust Soc Am*, vol. 101, no. 6, pp. 3602–3616, 1997.

[14] I.C. Bruce, F. Dinath, and T. J. Zeyl., "Insights into optimal phonemic compression from a computational model of the auditory periphery," in *Auditory Signal Processing in Hearing-Impaired Listeners, Int. Symposium on Audiological and Auditory Research (ISAAR)*, 2007, pp. 73–81.

[15] R. A. Ibrahim and I. C. Bruce, "Effects of peripheral tuning on the auditory nerves representation of speech envelope and temporal fine structure cues," in *The Neurophysiological Bases of Auditory Perception*, E. A. Lopez-Poveda el al., Ed., pp. 429–438. Springer NY, 2010.

[16] S. Liu and F. Zeng, "Temporal properties in clear speech perception," *J Acoust Soc Am*, vol. 120, no. 1, pp. 424–432, 2006.